Knowledge-Driven Mechanistic Enrichment of the Preeclampsia Ignorome

Columbia University Data Science Institute

Tiffany J. Callahan[1,2], Adrianne L. Stefanski[1], George Hripcsak[2], Lawrence E. Hunter[1]

[1]Computational Bioscience Program, University of Colorado AMC;
[2]Department of Biomedical Informatics, Columbia University

University of Colorado Anschutz Medical Campus

BACKGROUND

- Preeclampsia (PE) is life-threatening, acute-onset hypertension and proteinuria at ≥ 20 weeks gestation.[1] PE accounts for 40% of fetal mortality.[2] The only known cure is placenta delivery.
- Currently, more than one-third of all protein-coding genes have no known function or published literature (i.e., the ignorome).[3,4] Many disease-associated genes may provide important insight when examined within the context of of other diseases/phenotypes.
- An extensive body of scientific literature and data exist for PE. The PheKnowLator Ecosystem[5] helps users construct large-scale knowledge graphs (KGs) from a wide variety of biomedical data.

Can PheKnowLator be used to identify novel and actionable molecular mechanisms from the PE ignorome?

METHODS

The experimental design is highlighted in Figure 1. Data and scripts are available on GitHub: https://github.com/callahantiff/ignorenet.

Identification of the PE Molecular Signature

- A meta-analysis of domain expert-selected Gene Expression Omnibus (GEO)[6] studies.

Identification of Known PE-Associated Genes

- Literature-Driven. Mine PubTator[7], DisGeNET[8], and Malacards[9].
- Gene-Driven. Differentially expressed genes (DEGs) queried against PubAnntotation.

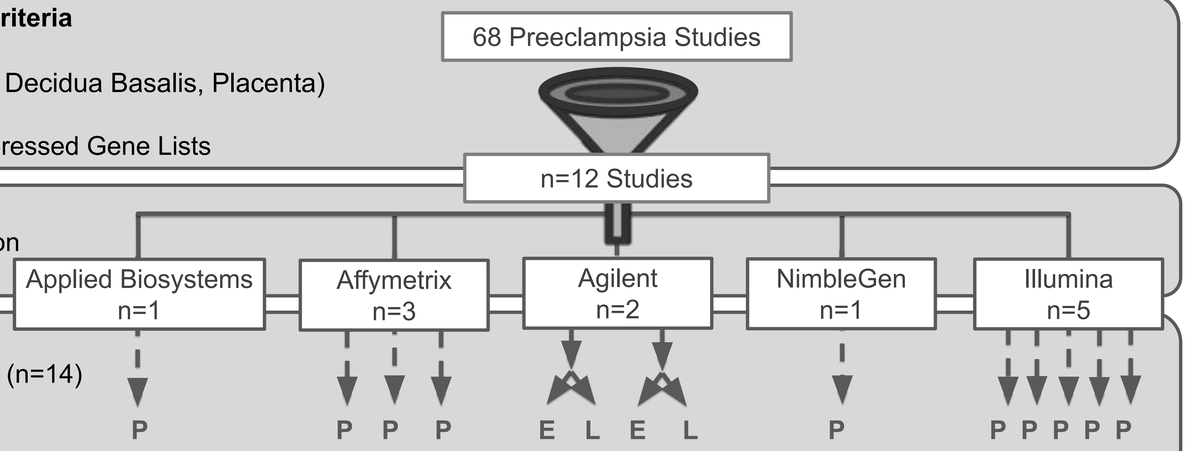
The PE ignorome was identified as genes from the PE molecular signature with no known PE-association in the literature.

PheKnowLator KG Enrichment

- Generate KG node embeddings using Walking RDF/OWL[10].
- The 100 nearest KG concepts (gold circles, Figure 1) to each ignorome gene were identified, reviewed by domain experts, and compared to gene set enrichment results produced by ToppGene[11].

PE GEO Microarray Data Selection Criteria

1. Human Gene Expression Data
2. Placenta Biopsy (i.e., Chorionic Villi, Decidua Basalis, Placenta)
3. Multi-Platform Microarray
4. Normalized Data or Differentially Expressed Gene Lists



Pre-Processing

- Normalization and Log2 Transformation
- Filter on First Quantile

Differential Expression

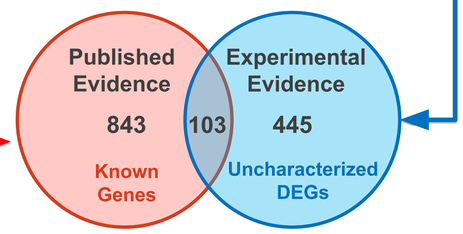
- LIMMA on Study Group Comparisons (n=14)
  - Control vs. Preeclampsia (P)
  - Control vs. Early Onset (E)
  - Control vs. Late Onset (L)

Differentially Expressed Genes ≥ 6 Study Comparisons: 548

Publication Annotations

Literature-Driven Approach	Gene-Driven Approach
Sources: PubTator, DisGeNET, Malacards	**Source:** PubAnnotation
Input: "Preeclampsia"; "HELLP Syndrome"; "Severe Preeclampsia"; "Placenta Disease"	**Input:** All differentially expressed genes and 18 preeclampsia-related identifiers
Output: 1,102 articles	**Output:** 1,962 articles

Unique Genes: 946



Annotation of Published and Uncharacterized PE Genes

Node embeddings derived from a PheKnowLator[7] KG

PheKnowLator Knowledge Graph Edges					(n=128,286 nodes)
Gene-Gene	594,100	Drug-Disease	1,216,900	Gene-Biological Process Gene-Cellular Component Gene-Molecular Function	265,002
Gene-Pathway	107,029	Disease-Phenotype	43,817		
Pathway-Disease	106,727	Gene-Disease	20,452	Pathway-Biological Process Pathway-Cellular Component Pathway-Molecular Function	17,906
Chemical-Pathway	711,043	Gene-Phenotype	120,288		

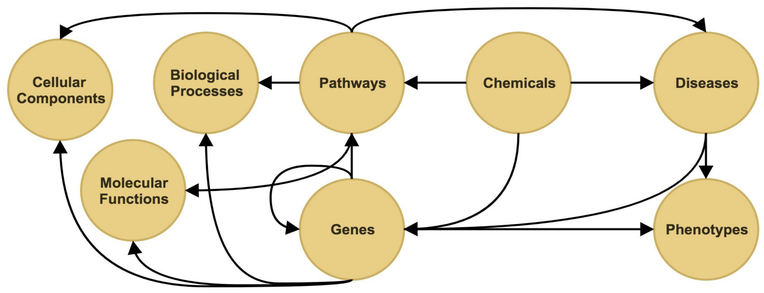


Figure 1. Overview of Experimental Approach for Finding the Preeclampsia Ignorome.

RESULTS

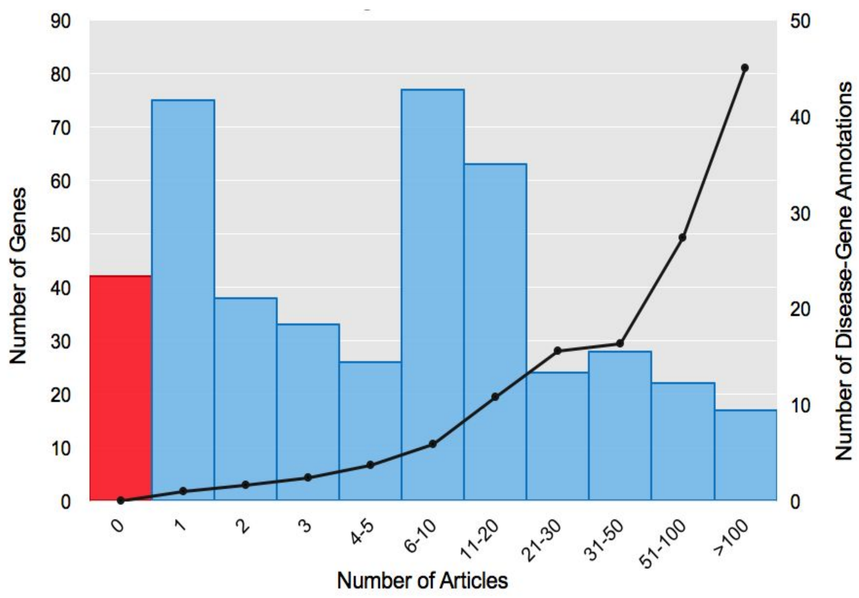


Figure 2. Illustrates the literature coverage of the 445 PE ignorome genes to other diseases. The x-axis represents the number of disease-annotated articles for each gene. The left y-axis shows the number of genes as bars. The right y-axis shows the number of diseases annotated to each PE gene.

- The PE ignorome contains 445 genes (Venn diagram, Figure 1).
- ToppGene Enrichment revealed that 90% of the PE ignorome genes were associated with a disease other than PE (Figure 2), most often neoplasms (48.7%).
- PheKnowLator-derived enrichment of the 100 KG concepts nearest to each PE ignorome gene resulted in 2,227 unique annotations.
- Expert reviewed reduced the 2,227 PheKnowLator annotations to 53 deemed worthy of experimental follow-up. None of the identified diseases, biological processes, cellular components, molecular functions, pathways, or phenotype associations overlapped with the ToppGene.
- Mechanistic explanations were derived for the 53 expert-selected annotations. An example of a novel disease association and mechanistic explanation is shown below:

TARDBP (NCBIGene:23435) - Amyotrophic Lateral Sclerosis (DOID:332)

TARDBP encodes the protein TDP-43 (PMID:28476168). AhR agonists increase TDP-43 in neurons. Placentas with high AhR expression during fetal development are highly susceptible to environmental toxicants (PMID:20354149). AhR has been proposed as a mechanism for the protective effects of cigarette smoke on preeclampsia (PMID:21864991).

CONCLUSIONS

- Expert-led multiplatform microarray meta-analysis and literature mining identified the PE ignorome (n=445). The majority of the ignorome genes were associated with a disease other than PE.
- The KG-based enrichment strategy produced 53 highly relevant novel PE associations, thus potentially identifying additional targets for prevention/intervention.
- The PheKnowLator Ecosystem can aid researchers and bench scientists in relevant and biologically-actionable discovery and provide new opportunities to leverage existing resources.

Limitations: Limited to transcriptionally-regulated molecules and should be considered with respect to the current lack of agreed upon standards for microarray meta-analysis.

References

1. Chaiworapongsa et al. *Nat Rev Nephrol* 10 (2014):466-80
2. Anderson et al. *Placenta* 33 (2012):S42-S47
3. Pandey et al. *PLoS One* 9 (2014):e88889
4. Riba et al. *Sci Rep* 6 (2016):24647
5. Callahan et al. *Zenodo* (2022):5716383
6. https://www.ncbi.nlm.nih.gov/geo/
7. https://www.ncbi.nlm.nih.gov/research/pubtator/
8. https://www.disgenet.org/
9. https://www.malacards.org/
10. https://github.com/bio-ontology-research-group/walking-rdf-and-owl
11. https://toppgene.cchmc.org/help/publications.jsp